

# Information Geometry and Sequential Monte Carlo Samplers

Aaron Sim<sup>a,\*</sup>, Sarah Filippi<sup>a</sup>, Michael P. H. Stumpf<sup>a,\*</sup>

<sup>a</sup>*Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences,  
Imperial College London, UK, SW7 2AZ*

---

## Abstract

This paper explores the application of methods from information geometry to the sequential Monte Carlo (SMC) sampler. In particular the Riemannian manifold Metropolis-adjusted Langevin algorithm (mMALA) is adapted for the transition kernels in SMC. Similar to its function in Markov chain Monte Carlo methods, the mMALA is a fully adaptable kernel which allows for efficient sampling of high-dimensional and highly correlated parameter spaces. We set up the theoretical framework for its use in SMC with a focus on the application to the problem of sequential Bayesian inference for dynamical systems as modelled by sets of ordinary differential equations. In addition, we argue that defining the sequence of distributions on geodesics optimises the effective sample sizes in the SMC run. We illustrate the application of the methodology by inferring the parameters of simulated Lotka-Volterra and Fitzhugh-Nagumo models. In particular we demonstrate that compared to employing a standard adaptive random walk kernel, the SMC sampler with an information geometric kernel design attains a higher level of statistical robustness in the inferred parameters of the dynamical systems.

**Keywords:** Information geometry, Sequential Monte Carlo, Bayesian inference, Dynamical Systems, mMALA

---

## 1. Introduction

Sequential Monte Carlo (SMC) techniques are very much the sampler *du jour* for a wide range of tasks traditionally tackled using Markov chain Monte Carlo (MCMC) methods [1, 2]. In the context of sequential Bayesian inference, for example, these include posterior sampling and model selection. The distinguishing characteristic of SMC sampling algorithms, as the name suggests, is a sequence of intermediate distributions  $\{\pi_1(x_1), \pi_2(x_2), \dots\}$ . In early applications of SMC methods (e.g. particle filters), such sequences would typically consist of non-homogeneous distributions defined on supports of different, often increasing, dimensions (i.e.  $\dim(\mathcal{X}_a) > \dim(\mathcal{X}_b)$  for  $a > b$ , where  $x_a \in \mathcal{X}_a$ ). Also, the algorithm outputs are usually samples from every distribution in the sequence. In more recent incarnations of SMC algorithms, in particular those viewed as alternatives to MCMC methods, these intermediate distributions only play a collective supporting role of bridging the gap between a tractable initial distribution and a target distribution of interest; it is also often the case, for example in tempered sequences, that the entire sequence is defined on a single support.

It is the presence of a common support which, together with the freedom of constructing the set of intermediate distributions, throws up an important design consideration for optimising the efficiency of the algorithm. The entire sequence, book-ended by the initial and target distributions, now defines a path in the space of distributions. As a proxy to the task of minimising the cumulative distances between the distributions in the chain, one seeks the ‘shortest’ continuous route between the initial and target distributions. This is not unlike the requirement in sequential importance sampling (SIS) for adjacent importance distributions to be ‘similar’ to each other. The overall aim of this paper is to explore two aspects of this design guidance. The first aspect is the identity of the sequence itself, while the second is the perturbation methods used to move along this pre-specified sequence of distributions. Apart

---

\*Corresponding authors

*Email addresses:* aaron.sim11@imperial.ac.uk (Aaron Sim), s.filippi@imperial.ac.uk (Sarah Filippi), m.stumpf@imperial.ac.uk (Michael P. H. Stumpf)

from several trivial examples where an analytical solution for the shortest path, or geodesic, exists (e.g. sampling from a multivariate normal distribution), consideration of this first aspect is usually restricted to the constrained problem of selecting the optimal spacing of distributions along a given, possibly arbitrarily chosen, path. For example, in a sequence of  $p$  tempered distributions

$$\{\pi_a(x) \sim \exp(\phi_a f(x))\}_{a \in \mathbb{T}}, \quad (1)$$

where  $\mathbb{T} = \{1, 2, \dots, p\}$ , and  $f$  some specified parametric function of  $x$ , selecting the optimal sequence simply involves tuning and selecting an optimal sequence of parameters  $\{\phi_a\}_{a \in \mathbb{T}}$ .

The specification of geodesics, together with the concept of movement in distribution-space strongly suggests that adopting a geometrical angle may provide a unifying framework for the consideration of the above design issues. Indeed such a geometrical framework was first introduced to MCMC in [3] where the authors employed ideas from information geometry to the design of efficient MCMC kernels and mooted its extension to population Monte Carlo (PMC) methods. Other recent examples of statistical applications of Riemannian geometry are its use in increasing the efficiency of Variational Bayesian methods [4] and in sensitivity analysis in stochastic models [5], amongst others.

Far from being an independent approach, SMC, together with other PMC methods, is often described as *parallel* MCMC, in part because the sampled variables, or particles, are typically perturbed along the sequences of intermediate distributions using MCMC proposals and are accepted or rejected based on the standard evaluation of the Metropolis-Hastings (MH) ratio. Seen in this context it is not surprising for theoretical developments in MCMC algorithms, specifically in the design of efficient kernel proposals, to find their way into SMC methods. In fact, as it has been demonstrated in [6], the well-documented advantage of PMC over MCMC samplers in addressing the issue of multiple distribution modes is preserved when adopting the differential geometric kernels. This paper follows this trend where we extend to the SMC sampler of Del Moral *et al* [2] the work of Girolami and Calderhead [3].

The elegance of a geometrical framework is very often the sole justification for its construction. However, in the case of the SMC sampler, the benefits extend beyond mere aesthetics; geometrically optimal placements of distributions and perturbation kernels will result in improved particle acceptance rates with greater effective sample sizes, which in turn lead to more robust and, hence, reliable sampling statistics over the course of the algorithm.

The rest of this paper is organised as follows. In section 2 we provide a brief review of the theoretical background to our work – namely SMC samplers and the relevant aspects of information geometry. In section 3 we consider geodesics on statistical manifolds and adapt the differential geometric MCMC kernels for use in SMC samplers. We illustrate the latter application to sequential Bayesian inference by way of three examples. We start with a trivial example involving the univariate Gaussian distribution and follow up with simulations involving the Fitzhugh-Nagumo and Lotka-Volterra ordinary differential equation (ODE) models. We conclude in chapter 4 with a summary and discussion.

*Note on notation.* We have adopted the Einstein summation convention. Also, components of a metric  $g$  and its inverse  $g^{-1}$  are written with covariant ( $g_{ij}$ ) and contravariant indices ( $g^{ij}$ ) respectively, such that  $g^{ik}g_{kj} = \delta_j^i$ , with  $\delta_j^i$  the Kronecker delta. Where there is no ambiguity, derivatives are abbreviated in the usual form as  $\partial_i \equiv \frac{\partial}{\partial \xi^i}$  for a given variable  $\xi^i$ . Component-wise vector multiplication which maps  $\mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  is specified by the  $*$  symbol, where  $(\mathbf{u} * \mathbf{v})_i \equiv u_i v_i$  (no sum), with  $u_i$  and  $v_i$  the  $i$ th components of vectors  $\mathbf{u}$  and  $\mathbf{v}$  respectively.

## 2. Theoretical background

### 2.1. Sequential Monte Carlo Samplers

There are, at present, several different implementations of SMC algorithms (see, for instance, [1, 7]). With regards to the ability to admit intermediate distributions defined on a common space – a necessary requirement of our attempt to build a geometrical framework – and the flexibility to define general moves between populations, it is the SMC sampler implementation by Del Moral *et al* [2] which proves the most natural. We adopt this version in this paper and, in accordance with its authors, refer to it throughout simply as the *SMC sampler*.

Similar to SIS, the SMC sampler is used to sample successively from a sequence of distributions. Unlike SIS, however, the SMC sampler does not require one to calculate the intermediate importance distributions. This is achieved through the introduction of artificial backward-in-time Markov kernels  $L_a$  alongside the usual forward-in-time versions  $K_a$ , the effect of which is a reduction of the complexity from  $O(N^2)$  to  $O(N)$ , where  $N$  is the number of sampled particles.

Let  $\pi_p$ ,  $p \in \mathbb{Z}$ , be a target distribution for sampling. One specifies a preceding sequence of distributions  $\{\pi_1, \pi_2, \dots, \pi_{p-1}\}$ , with all  $\{\pi_a\}_{a \in \mathbb{T}}$  defined on the support of  $\pi_p$ . Let  $\{\eta_a\}_{a \in \mathbb{T}}$  represent the sequence of importance distributions where only the initial importance distribution  $\eta_1$  is explicitly specified; the most natural choice is  $\eta_1 = \pi_1$ . In practice, the normalisation constants are unknown and one works instead with the unnormalised sequence  $\{\gamma_a\}_{a \in \mathbb{T}}$ , where

$$\pi_a = \frac{\gamma_a}{Z_a}, \quad Z_a \in \mathbb{R}. \quad (2)$$

One samples a population of  $N$  particles  $\{\xi_a^{(n)}\}_{n \in \{1, 2, \dots, N\}}$  from the importance distribution  $\eta_a$  by perturbing particles from the previous population via the transition kernels  $K_a(\xi_a^{(n)} | \xi_{a-1}^{(n)})$  – e.g. local random walks, Gibbs moves, etc. In this paper we use MCMC kernels for these local moves, the advantage being that the guaranteed convergence to each  $\pi_a$  allows for several simplifying approximations to be made in the algorithm (see [2]). In particular, one has a simple approximation of a near-optimal expression of  $L_a$  in terms of  $K_a$  and  $\pi_a$ , which, in turn, simplifies the sequential updating of the particle weights – i.e. the incremental factor to update the weights of a particle  $\xi_{a-1}^{(n)}$  is

$$\tilde{w}_a(\xi_{a-1}^{(n)}, \xi_a^{(n)}) = \frac{\gamma_a(\xi_{a-1}^{(n)})}{\gamma_{a-1}(\xi_{a-1}^{(n)})}. \quad (3)$$

The SMC sampler is summarised below (Algorithm 1). For more details of the construction of the algorithm and proofs of consistency, we refer the reader to [2].

---

**Algorithm 1:** The SMC sampler with MCMC kernels

---

**Input:** No. of particles per population  $N$ , Sequence of distributions  $\pi_a$ ,  $a = 1, \dots, p$ , Effective sample size (ESS) threshold  $T$ .

**Output:** Sampled particles  $\xi_p^i$ , Particle weights  $W_p^i$ ,  $i = 1, \dots, N$ .

- 1 Initialise  $a = 1$ ,  $\text{ESS} = N$ ;
- 2 Sample particles  $\xi_1^{(n)}$  from prior  $\pi_1$ ;
- 3 Set weights  $W_1^{(n)} = \frac{1}{N}$ ;
- 4 **for**  $a \leq p$  **do**
- 5     **if**  $\text{ESS} < T$  **then**
- 6         Resample particles  $\{\xi_{a-1}^{(n)}\}$  from weighted multinomial distribution  $\{(\xi_{a-1}^{(n)}, W_{a-1}^{(n)})\}$ ;
- 7         Reset weights  $W_{a-1}^{(n)} = \frac{1}{N}$ ,  $\forall i = 1, \dots, N$ ;
- 8     **end**
- 9     **for**  $n = 1, 2, \dots, N$  **do**
- 10         Draw  $\xi_a^{(n)} \sim K_a(\cdot | \xi_{a-1}^{(n)})$ , where  $K_a$  is a MCMC kernel;
- 11         Evaluate incremental weight  $\tilde{w}_a(\xi_{a-1}^{(n)}, \xi_a^{(n)}) = \frac{\gamma_a(\xi_{a-1}^{(n)})}{\gamma_{a-1}(\xi_{a-1}^{(n)})}$ ;
- 12         Update particle weight  $\tilde{W}_a^n = W_{a-1}^n \cdot \tilde{w}_a(\xi_{a-1}^{(n)}, \xi_a^{(n)})$ ;
- 13     **end**
- 14     Normalise particle weights  $W_a^n = \tilde{W}_a^n / \sum_{m=1}^N \tilde{W}_a^m$ ;
- 15     Calculate  $\text{ESS} = 1 / \sum_{n=1}^N |W_a^n|^2$ ;
- 16     Set  $a = a + 1$ ;
- 17 **end**
- 18 Return  $\{\xi_p^{(n)}, W_p^n\}_{i \in \{1, \dots, N\}}$ .

---

For the remainder of the paper, where there is no ambiguity, we drop the particle index and simply write  $\xi_a^{(n)} \rightarrow \xi_a$ .

## 2.2. Information geometry

The premise of all of information geometry is the observation that the space of probability distributions on a set  $\mathcal{X}$  can be viewed as a Riemannian manifold with a unique metric  $g$  [8]. Clearly, this is an infinite-dimensional manifold. In most applications, however, one considers the projection to a finite-dimensional submanifold  $\mathcal{S}$  by restricting to a specific statistical model of choice as represented by a set  $\Xi$  of parameters. By expressing the set of coordinate functions on the manifold in terms of model parameters  $\xi \in \Xi$ , it is easy to conceptualise the manifold structure of  $\mathcal{S}$ . For example, the space of two-dimensional Gaussian distributions  $\mathcal{N}(\mu, \Sigma)$  is a five-dimensional manifold with (non-canonical) coordinate functions  $\{\mu_1, \mu_2, \Sigma_{11}, \Sigma_{12}, \Sigma_{22}\}$ .

The coordinate-specific metric  $g_{ij}(\xi)$  on  $\mathcal{S} = \{p(x; \xi) \mid \xi \in \Xi\}$  is defined by the expected Fisher information matrix

$$g_{ij}(\xi) := E_\xi[\partial_i \ell_\xi \partial_j \ell_\xi] = \int \partial_i \ell(x; \xi) \partial_j \ell(x; \xi) p(x; \xi) dx, \quad (4)$$

where  $\ell(x; \xi) \equiv \log p(x; \xi)$ , and  $E_\xi$  is the expectation with respect to the probability distribution  $p(x; \xi)$ . Using the property  $\int p(x; \xi) dx = 1$ , one derives the useful alternative form of the Fisher metric

$$g_{ij}(\xi) = -E_\xi[\partial_i \partial_j \ell_\xi]. \quad (5)$$

Apart from the invariance of  $g_{ij}$  under a sufficient statistic map, the foundations of information geometry contain few other constraints. Specifically there are no restrictions on the class of statistical models or on the identity of the measure space  $\mathcal{X}$ , which can be discrete or continuous, thereby explaining its wide applicability across many different fields. For example, in systems biology contexts, we often have

$$\mathcal{X} = (\mathbb{R}^+)^{T_1} \times (\mathbb{R}^+)^{T_2} \times \dots \times (\mathbb{R}^+)^{T_S}, \quad (6)$$

where each sample  $x \in \mathcal{X}$  represents a set of non-negative measurements (species abundance, chemical concentration, etc) of  $S$  separate biological entities, each measured at  $T_s$  separate time-points ( $s = 1, \dots, S$ ).

With the metric  $g_{ij}$ , one proceeds to import the concepts and structures of Riemannian geometry like distances, geodesics, curvature, etc to a wide range of statistical analyses (see [8] for a concise introduction to the relevant aspects of Riemannian geometry). For example, we can define parallel transport on the space of distributions using the Levi-Civita connection  $\nabla_{\partial_i} \partial_j = \Gamma^k_{ij} \partial_k$ , where the Christoffel symbols  $\Gamma$  are expressed, in the usual way, in terms of the Fisher metric and its derivatives as

$$\Gamma^k_{ij} = -\frac{1}{2} g^{kl} (\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}). \quad (7)$$

Using this expression we have the geodesics  $\gamma : \mathbb{R} \rightarrow \mathcal{S}$  on  $\mathcal{S}$  as solutions to the geodesic equation

$$\ddot{\gamma}^k(t) + \dot{\gamma}^i(t) \dot{\gamma}^j(t) (\Gamma^k_{ij})_{\gamma(t)} = 0, \quad (8)$$

where the dot refers to differentiation w.r.t.  $t \in \mathbb{R}$ , and the indices indicate the  $i$ th component in the given coordinate parameterisation, i.e.  $\gamma^i(t) := \xi^i(\gamma(t))$ , etc.

In information geometry, the infinitesimal distance between two distributions is directly related to the Kullback-Leibler (KL) divergence  $KL(\cdot \mid \cdot)$ . With  $ds^2 = g_{ij} d\xi^i d\xi^j$ , we have [9]

$$KL(p(x; \xi + d\xi) \mid p(x; \xi)) = \frac{1}{2} ds^2. \quad (9)$$

## 3. Information geometric design of SMC samplers

### 3.1. Geodesics on statistical manifolds

In many implementations of the SMC sampler, the sampling of each intermediate distribution is optimised by minimising the KL-divergence from the previous distribution in the sequence. This is often represented symbolically as  $\pi_a \approx \pi_{a+1}$ , where  $a = 1, 2, \dots, p$ . Together with the relation in (9), this suggests that, in the asymptotic limit  $p \rightarrow \infty$ , the full SMC sampler is optimised by selecting the

intermediate distributions to lie on the geodesic connecting the fixed initial and final target distribution boundary points. We illustrate this by way of the following trivial example.

*Sampling from a univariate Gaussian distribution.* Assuming that one has a process of sampling from the standard normal  $\mathcal{N}(0, 1)$ , one can use the SMC sampler to sample from an arbitrary distribution  $\mathcal{N}(\mu', \sigma'^2)$ . We consider sequences of 25 distributions on each of three separate paths on the univariate Gaussian manifold<sup>1</sup>, with one of the paths being the geodesic. The analytic solution to the geodesic equation (8) with fixed boundary points is provided in Appendix A.1. We compare the optimality of the SMC sampler by tracking the effective sample size as the SMC sampler progresses through the intermediate distributions.

In the coordinate representation  $(\mu, \sigma^2)$ , we select arbitrary initial and target distributions  $p_1 = (0, 1)$  and  $p_{25} = (5, 3)$  respectively. We employ a local random walk kernel with a fixed uniform proposal. Now in the SMC sampler algorithm (Algorithm 1), the expression for the incremental weights  $\tilde{w}_a$  is a good approximation for MCMC kernels only if the successive populations are close, i.e.  $\pi_{a-1} \approx \pi_a$  (see [2] eq. 31). However, given that the effect of altering the distances between distributions is what we are aiming to test, there may be instances where this assumption is no longer valid. In its place, therefore, we use the full expression ([2] eq. 26) and particle approximation

$$\tilde{w}_a(\xi_{a-1}, \xi_a) = \frac{\gamma_a(\xi_a)}{\int_{\text{all } \zeta} \gamma_{a-1}(\zeta) K_a(\xi_a | \zeta)} \approx \frac{\gamma_a(\xi_a)}{\sum_{n=1}^N W_{a-1}^n K_a(\xi_a^{(n)} | \xi_{a-1}^{(n)})}, \quad (10)$$

where the full analytic expression for the kernel density  $K_a(\xi_a^{(n)} | \xi_{a-1}^{(n)})$  is given in Appendix A.2.

Without resampling, we track the average progression of the mean effective sample sizes (ESS) across ten independent iterations of the SMC sampler. The results, together with the paths on the manifold, are shown in Figure 1. It is clear that selecting the distributions to lie on the geodesics leads to a slower decline of the ESS over the course of the SMC sampler.

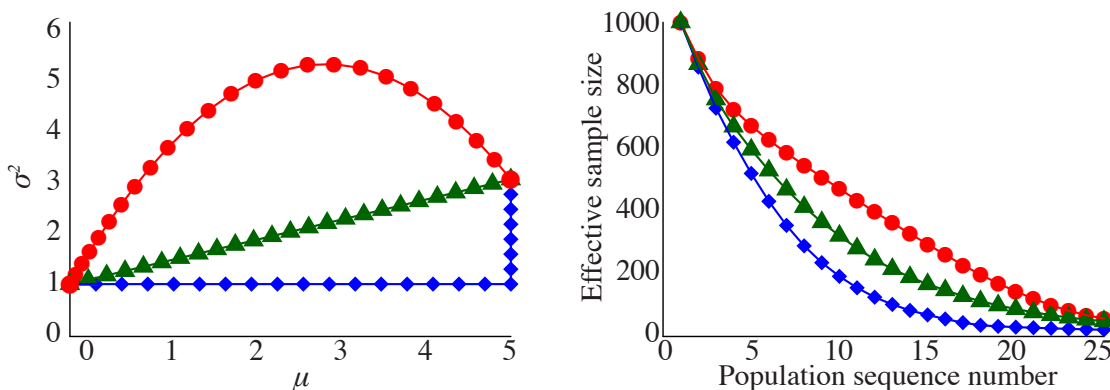


Figure 1: **Left:** Sample paths between initial distribution  $\mathcal{N}(0, 1)$  and the target distribution  $\mathcal{N}(5, 3)$ , each with 25 intermediate distributions (including end-distributions). **Right:** Evolution of ESS over 25 populations. The circular points (red) represent the distributions on the geodesic between the fixed boundaries; the triangles (green) indicate the naive straight-line path which implicitly (and erroneously) assumes a Euclidean metric; the diamonds (blue) indicate a two-staged path with the  $\mu$  coordinate the first to be fully adjusted.

In practice, one would simply transform the samples from the standard normal  $x \rightarrow x\sigma' + \mu'$ . However, the availability of analytical solutions to the geodesic equation for Gaussian distributions presents an opportunity to employ the SMC sampler and to observe the effect of locating distributions on geodesics. Unfortunately, such analytical solutions are almost always unavailable. Indeed, as mentioned in the introduction, the procedure of writing down a tempered sequence (1) simply side-steps the issue by adopting the possibly non-optimal solution of fixing all but one of the parameters to the target distribu-

<sup>1</sup>Note that no attempt was made, on either geodesic or non-geodesic paths, to ensure equal Fisher distance between all pairs  $\int_a^{a+1} ds$ , as was done in [10]. Should we choose to space the distributions evenly, we fully expect the relative performance of the three paths, and our conclusions, to remain unchanged.

tion values. Effectively, this is equivalent to taking a one-dimensional submanifold, with parametrising coordinate  $\phi$ , in place of the true geodesic, and then attempting to atone for our sins by selecting a sequence of  $\{\phi_a\}_{a \in \mathbb{T}}$  which minimises the cumulative distances for that particular path. One can also adopt a brute force approach to reducing the distance between each intermediate distribution by simply selecting a large number of intermediate distributions.

### 3.2. Differential geometric MCMC kernels for SMC transitions

A non-optimal placement of intermediate distributions in SMC can be mitigated somewhat by employing an efficient kernel for transitions between distributions. In this section we describe one such method – the manifold Metropolis-adjusted Langevin algorithm (mMALA) [3]. mMALA is an algorithm which prescribes, using the local geometry of the statistical manifold, the discretised flow of a particle on a given space, whereby the sampled positions of the particle over time follow a specified distribution.

The mMALA MCMC kernel combines a discretised Brownian motion with drift proposal with the standard Metropolis-Hasting acceptance step. Let  $\xi \in \mathbb{R}^D$  be a random vector parametrising distributions  $p(x; \xi)$ ,  $x \in \mathcal{X}$ . As described above, we treat  $\xi$  as the coordinate functions of a  $D$ -dimensional manifold  $\mathcal{S}$ . Define  $\tilde{p}(\xi)$  to be a density<sup>2</sup> on  $\mathcal{S}$ . In the special case where the target distribution is proportional to the likelihood of the set of  $S$  observed samples  $\{x_s\}$  from  $\mathcal{X}$ , i.e.  $\tilde{p}(\xi) \propto \prod_{s=1}^S p(x_s; \xi)$ , one can calculate the Fisher information matrix  $g$  and track the discretised Langevin diffusion on the Riemannian manifold  $(\mathcal{S}, g)$ . As shown in [3] The proposal density is the Gaussian

$$q(\xi_{\tau+1} \mid \xi_\tau) \sim \mathcal{N}(\mu(\xi_\tau, \epsilon), \epsilon^2 g^{-1}), \quad (11)$$

with  $\epsilon$  the Langevin diffusion discretisation step size,  $\tau$  the discrete time index, and where the components of the deterministic drift are given by

$$\mu^i(\xi_\tau, \epsilon) = (\xi_\tau)^i + \frac{\epsilon^2}{2} (g^{ij} (\partial_j \ell(\xi_\tau))) - \epsilon^2 (g^{ik} (\partial_j g_{kl}) g^{lj}) + \frac{\epsilon^2}{2} (g^{ij} g^{kl} (\partial_j g_{kl})), \quad (12)$$

where  $\ell(\xi) \equiv \log \tilde{p}(\xi)$ .

The geometric design of efficient transition kernels for SMC is easily adapted from the MCMC context. Consider a sequence  $\{\tilde{p}_a(\xi)\}_{a \in \mathbb{T}}$ , defined on a common space  $\mathcal{S}$ . Where in MCMC a single metric  $g_{ij}(\xi)$  is defined from the MCMC invariant density, here we have a sequence of metrics  $\{(g_a)_{ij}\}_{a \in \mathbb{T}}$ . Concomitantly, the set of proposal densities (11) is replaced by a sequence of proposal densities

$$\left\{ q_a(\xi_{a+1} \mid \xi_a) = \mathcal{N}(\mu_a(\xi_a, \epsilon), \epsilon^2 g_a^{-1}) \right\}_{a \in \mathbb{T}}, \quad (13)$$

where  $\mu_a(\xi_a, \epsilon)$  is given by (12) with  $g$  replaced, in sequence, by the metrics in  $\{g_a\}_{a \in \mathbb{T}}$ .

Assuming that  $\tilde{p}_a \approx \tilde{p}_{a+1}$  we expect a relatively high acceptance rate in the MH step. We can, therefore, get away with performing relatively fewer MCMC iterations (perhaps even just one per distribution in the sequence). Although this reduces the overall computational cost of the algorithm, it is dependent on the quality of the mixing property of the kernel, the latter of which can be improved, albeit at the expense of a decrease in acceptance rate, by choosing a larger Langevin diffusion step size  $\epsilon$ .

Apart from the choice of the initial distribution, there are two key algorithmic parameters to set: the discretisation step size,  $\epsilon$ , and the number of populations,  $p$ . We demonstrate the application of the mMALA—SMC combination and observe the implications of various parameter combination choices by way of the following simple example.

### 3.3. Example: Univariate Gaussian parameter inference

Consider  $S$  observations,  $x_i$ , drawn from a univariate normal distribution  $\mathcal{N}(\mu, \sigma^2)$  with parameters  $\mu$  and  $\sigma$  to be inferred using the SMC sampler. Adopting  $\xi = (\xi_1, \xi_2) = (\mu, \sigma)$  as the coordinate functions

---

<sup>2</sup>Note that this density is an additional structure on the manifold;  $\tilde{p}(\xi)$  is defined on  $\mathcal{S}$ , whereas for  $s \in \mathcal{S}$ ,  $p(x; \xi(s))$  is defined on  $\mathcal{X}$ .

of a manifold  $\mathcal{S}$ , we assume the normal priors  $\pi_1(\xi_1) = \mathcal{N}(u_1, v_1^2)$  and  $\pi_2(\xi_2) = \mathcal{N}(u_2, v_2^2)$  for some  $u_1, v_1, u_2, v_2 \in \mathbb{R}$ . We define a sequence of  $p$  tempered distributions [11]  $\{\tilde{p}_a(\xi)\}_{a \in \mathbb{T}}$ ,  $\mathbb{T} = \{1, \dots, p\}$ , where

$$\tilde{p}_a(\xi) = \pi(\xi) \left( \prod_{s=1}^S \frac{1}{\sqrt{2\pi\xi_s^2}} \exp - \frac{(x_s - \xi_1)^2}{2\xi_s^2} \right)^{\phi_a}, \quad (14)$$

with  $0 = \phi_1 < \phi_2 < \dots < \phi_{p-1} < \phi_p = 1$ , and  $\pi(\xi) = \text{diag}(\pi_1(\xi_1), \pi_2(\xi_2))$ . It can be shown that for tempered distributions of the exponential form (e.g. (14)), the sum of symmetrised Kullback-Leibler divergences between the distributions is minimised by adopting a geometric tempering sequence where  $\phi_{a+1}/\phi_a = \text{const}$  [12]. Therefore, in the examples that follow, we adopt the geometric sequence with

$$\phi_a = \phi_2^{-\frac{a-2}{p-2}+1}, \quad \text{and} \quad \phi_1 = 0. \quad (15)$$

In addition, in this example, we fix  $\phi_2 = 5 \times 10^{-4}$ .

Using (5) the metric for each distribution in the sequence is given by

$$g_a(\xi) = \begin{pmatrix} \frac{1}{v_1^2} + \frac{S\phi_a}{\xi_2^2} & 0 \\ 0 & \frac{1}{v_2^2} + \frac{2S\phi_a}{\xi_2^2} \end{pmatrix}. \quad (16)$$

For step-size  $\epsilon$ , the mMALA drift and the covariance matrix of the diffusion term on  $\mathcal{S}$  are, respectively,

$$\begin{pmatrix} \mu_\mu \\ \mu_\sigma \end{pmatrix} = \frac{\epsilon^2}{2} \begin{pmatrix} \frac{v_1^2 \xi_2^2}{C_1} \left[ -\frac{\xi_1 - u_1}{v_1^2} + \sum_{s=1}^S \frac{\phi_a(x_s - \xi_1)}{\xi_2^2} \right] \\ \frac{v_2^2 \xi_2^2}{C_2} \left[ -\frac{\xi_2 - u_2}{v_2^2} - \frac{S\phi_1}{\xi_2} + \sum_{s=1}^S \frac{\phi_a(x_s - \xi_1)^2}{\xi_2^3} + \frac{2S\phi_a}{\xi_2} \left( \frac{v_2^2}{C_2} - \frac{v_1^2}{2C_1} \right) \right] \end{pmatrix}, \quad (17)$$

$$\Sigma_{ij} \equiv \epsilon^2 g^{ij} = \epsilon^2 \begin{pmatrix} \frac{v_1^2 \xi_2^2}{C_1} & 0 \\ 0 & \frac{v_2^2 \xi_2^2}{C_2} \end{pmatrix}, \quad (18)$$

with the  $C_1 \equiv \xi_2^2 + v_1^2 S \phi_a$  and  $C_2 \equiv \xi_2^2 + 2v_2^2 S \phi_a$ . An important role played by the prior scale parameters  $v_1$  and  $v_2$  is that of setting a lower bound on the components of the metric  $g_a$ . The implications of this effect are examined in Section 3.4.

*Parameter inference.* We sample 60 points from the distribution  $\mathcal{N}(\mu', \sigma'^2) = \mathcal{N}(50, 10^2)$  and let the prior distribution be centred on  $(\mu', \sigma')$ , i.e.  $u_1 = 50$  and  $u_2 = 10$ , with  $v_1 = 20, v_2 = 2.5$ . Selecting a sequence of 45 populations and using 1500 particles, we run the SMC sampler with  $\epsilon = 0.4$  and resampling fraction  $T = 0.3$ . The estimates of the joint and marginal posteriors of  $\mu, \sigma$  are presented in Figure 2. As expected, the SMC sampler with mMALA transitions lead to a posterior which is centred around  $(\mu, \sigma) = (50, 10)$ .

*mMALA drift.* In order to visualise the path of the particles and to understand the effects of varying the parameters  $p$  and  $\epsilon$  on the effectiveness of the SMC sampler, we focus on the deterministic perturbation at each intermediate stage by setting the diffusion term to zero and accepting all particles without resampling. We select 24 points on the manifold  $\mathcal{S}$  from a grid surrounding the sample and observe the drift (12) for varying numbers of intermediate populations  $p$  and step-sizes  $\epsilon$ . The plots of the paths in  $\mathcal{S}$  are given in Figure 3. We can already deduce several features of the mMALA drift. For a fixed step-size,  $\epsilon$ , a minimum acceptable number of populations for the SMC run with  $p_{\min} \sim 45$ . If  $p < 45$ , the drift is too weak and the kernels do not mix well; if  $p \geq 45$ , the particles do indeed drift toward the target coordinates with the variance in particle end-points decreasing as  $p$  increases. If  $\epsilon$  is too small, the drift is too weak, resulting in a poor performance of the SMC sampler; if  $\epsilon$  is too large, the sharp changes in drift direction is an indication of the expected breakdown in the first-order Euler discretisation of the Langevin diffusion. The implication of this last point for SMC samplers is that a higher proportion of particles will be rejected in the MH-step leading to a lower number of effective particles used in the sampler. These results support the intuitive expectation that there is not just a minimum step size and total number of perturbation steps for effective sampling, but also a negative impact on the efficiency of the sampler when these tuning parameters are too large.

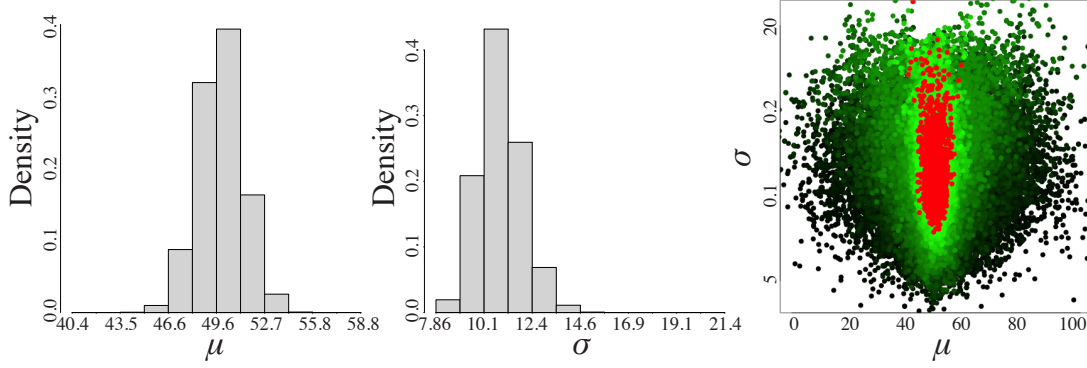


Figure 2: **Left/middle:** Weighted histogram of particles representing the estimated marginal distributions of the mean and standard deviation parameters. **Right:** Scatterplot of particles over 45 populations for a simulated univariate model  $\mathcal{N}(50, 10^2)$ . The initial particles are sampled from the prior  $\pi(\xi) \equiv p_0(\xi) \sim \text{diag}(\mathcal{N}(50, 20^2), \mathcal{N}(10, 2.5^2))$  and are coloured black in the scatter plot. The particles of the intermediate populations are represented in increasingly brighter shades of green with particles of the final population, representing the posterior, coloured red.

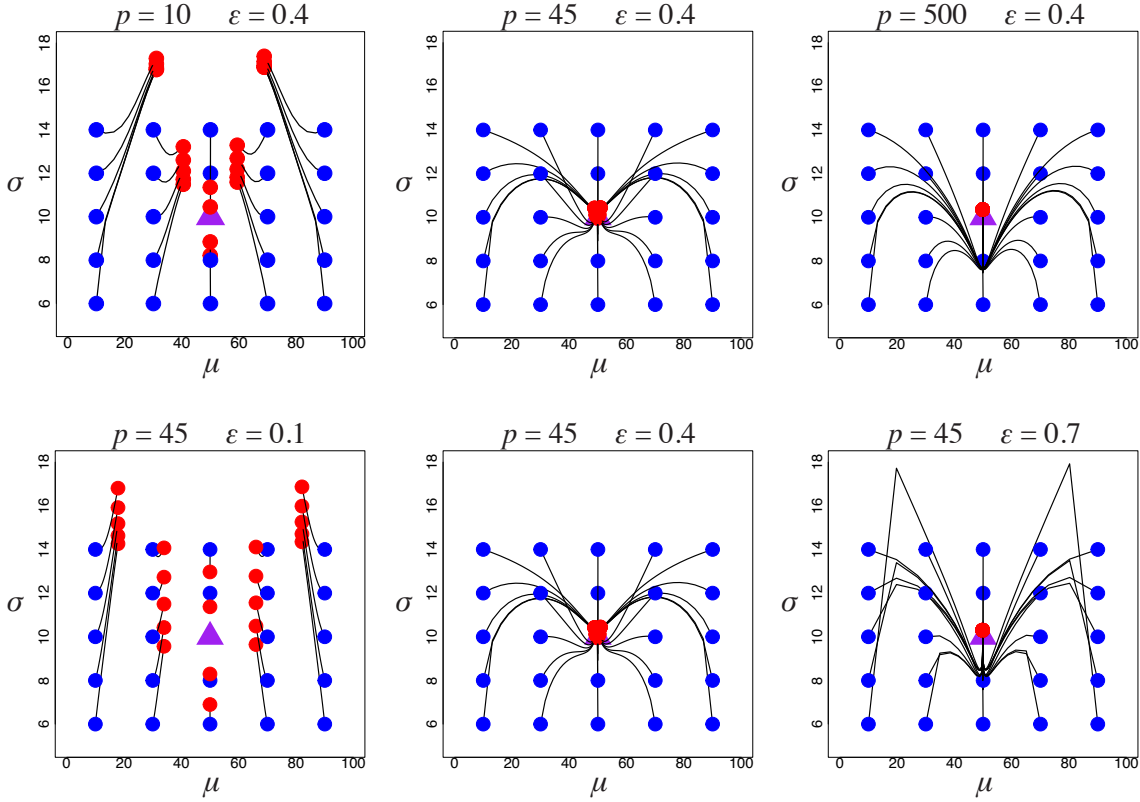


Figure 3: mMALA particle drift paths for the univariate normal model for different numbers of intermediate populations  $p$  (top row), and step sizes  $\epsilon$  (bottom row). The three population sizes and step sizes simulated are  $p = 10, 45, 500$ , and  $\epsilon = 0.1, 0.4, 0.7$  respectively. The **purple** triangle marks the simulated mean and standard deviation  $(\mu', \sigma') = (50, 10)$ ; the **blue** points represent the initial 24 particle samples and the **red** points the perturbed particles at the final population.

The results from this simple example hints at a strategy for optimising the differential geometric kernels of the SMC sampler: keeping constant the product of the square of the step-size and population number, i.e.

$$\epsilon^2 p \sim \text{const}, \quad (19)$$

we seek to minimise the number of populations whilst ensuring that the step-size is not so large that the precision of the algorithm is impacted by an increased rejection rate. For example, we found that the drift behaviour with  $\epsilon = 0.4$ ,  $p = 45$  is essentially identical to the set-up with  $\epsilon = 0.2$ , and  $p = 180$  ( $0.4^2 \times 45 = 0.2^2 \times 180$ ). Without factoring in the particle acceptance rate, the efficiency of the algorithm



simply scales with the number of populations.

Other than the tuneable parameters,  $p$  and  $\epsilon$ , it is the identity of the starting distribution, e.g. the prior in the context of Sequential Bayesian inference, which has a big impact on the viability of the mMALA transitions for SMC samplers. We defer this discussion to the following section where the implications are particularly significant.

### 3.4. Application to parameter inference in dynamical systems

Following [3], we consider a system of differential equations

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}, \xi, t), \quad (20)$$

where  $\mathbf{x}$  is a  $D$ -dimensional vector,  $\xi$  the model parameter vector, and  $t$  the time variable. The observations  $Y$  of  $\mathbf{x}$  are made of the underlying state at  $\tau$  time-points  $\{t_1, t_2, \dots, t_\tau\}$  with a known observation noise model, i.e.

$$Y = X + E, \quad (21)$$

where  $X = X(\mathbf{x}_0, \xi) = (\mathbf{x}(t_1) | \mathbf{x}(t_2) | \dots | \mathbf{x}(t_\tau))^T$  is a solution to (20), and  $E$  the observation noise.  $Y$ ,  $X$ , and  $E$  are  $(\tau \times D)$ -dimensional matrices. For given initial conditions  $\mathbf{x}_0$  and prior  $\pi(\xi)$ , the task is to estimate the posterior

$$p(\xi | Y, \mathbf{x}_0, E) \propto \pi(\xi) \cdot L(Y | \xi, \mathbf{x}_0, E), \quad (22)$$

where  $L(\cdot | \cdot)$  is the likelihood function. For example, as described in [3], for a noise model given by a time-independent normal distribution with variance  $\sigma_d^2$ ,  $d \in \{1, \dots, D\}$ , we have

$$L(Y | \xi, \mathbf{x}_0, E) = \prod_d^D \mathcal{N}(X(\xi, \mathbf{x}_0)_{\cdot, d}, \Sigma_d), \quad (23)$$

where  $\Sigma_d = \mathbb{I}_\tau \sigma_d^2$ , and  $X_{\cdot, d}$  denotes the time-series vector for species  $d$ . For a log-normal noise model with corresponding parameters, we have

$$L(Y | \xi, \mathbf{x}_0, E) = \prod_d^D \log \mathcal{N}(X(\xi, \mathbf{x}_0)_{\cdot, d}, \Sigma_d). \quad (24)$$

Observation noise in realistic dynamical systems are, however, likely to be more complicated. The only consequence of deviating from the above vanilla noise models is an added complexity in the Fisher metric calculations. As an example of the additional algebra involved, we have examined the implications of two such modifications – heteroscedasticity and a truncation of the normal noise model. The details of these Fisher metric calculations are collected in Appendix C.1. Nevertheless, since there are no conceptual novelties arising from cases with such adjustments, we have simply adopted the normal model of (23) for the remainder of this section.

The problem can now be given a geometrical construction. We treat the model parameters  $\xi$  as coordinate functions on a manifold  $\mathcal{S}$ . The combination of the differential equations (20) and an observation noise model allows us to associate to each point in  $\mathcal{S}$  a probability distribution as follows. Define the measure space

$$\mathcal{X} \cong \underbrace{(\mathbb{R}^\tau \times \mathbb{R}^\tau \times \dots \times \mathbb{R}^\tau)}_{D \text{ terms}}. \quad (25)$$

The density  $p(Y_{\cdot, n}; \xi)$  is given by the posterior in (22), i.e.

$$p(Y_{\cdot, n}; \xi) = \kappa \pi(\xi) \cdot L(Y_{\cdot, n} | \xi, \mathbf{x}_0, E), \quad (26)$$

where  $\kappa \in \mathbb{R}$  is a constant. Using this density we can then proceed to define the Fisher metric  $g_{ij}(\xi)$  on  $\mathcal{S}$  via (5). For both the Gaussian and log-normal noise model, the likelihood has an exponential form.

Hence, abbreviating  $p(x; \xi) \equiv \kappa \pi(\xi) \exp \Phi(\xi)$ , we have

$$\begin{aligned} -\partial_i \partial_j (\log p(Y_{\cdot, n}; \xi)) &= -\partial_i \partial_j \Phi - \frac{(\partial_i \partial_j \pi(\xi))}{\pi(\xi)} + \frac{(\partial_i \pi(\xi))(\partial_j \pi(\xi))}{\pi^2(\xi)} \\ &\equiv -\partial_i \partial_j \Phi + h_{ij}(\xi), \end{aligned} \quad (27)$$

where  $h_{ij}(\xi)$  depends only on  $\xi$  via the prior  $\pi(\xi)$  and not on  $Y_{\cdot, n}$ . We evaluate  $h_{ij}(\xi)$  for several typical priors – uniform, multivariate normal (MVN), and the component-wise (CW) log-normal – as<sup>3</sup>

$$\text{Uniform:} \quad h_{ij}(\xi) = 0, \quad (28)$$

$$\text{MVN:} \quad h_{ij}(\xi) = (\Sigma^{-1})_{ij}, \quad (29)$$

$$\text{CW log-normal:} \quad h_{ij}(\xi) = \frac{\delta_{ij}}{(\xi^i \sigma^i)^2} (1 - \log \xi^i + \mu^i), \quad (30)$$

where the  $\mu_i$  and  $\Sigma_{ij}$  are usual components of the mean vector and covariance matrix in  $\dim(\mathcal{S})$ -dimensions; we also assume that for the uniform prior,  $h_{ij}(\xi)$  is evaluated away from the boundary of the non-zero support where  $\partial_i \pi(\xi)$  is undefined.

In the framework of SMC we now define, on the shared measure space (25), the sequence of  $p$  distributions

$$p_a(x; \xi) = \kappa \pi(\xi) \cdot [L(Y | \xi, \mathbf{x}_0, E)]^{\phi_a}, \quad (31)$$

with  $a \in \mathbb{T}$  and  $0 = \phi_1 < \phi_2 < \dots < \phi_p = 1$ . From this sequence of distributions we have a matching sequence of Riemannian manifolds  $\{(\mathcal{S}_a, g_a)\}_{a \in \mathbb{T}}$ . Similar to the evaluation in [3], the Fisher metrics  $(g_a)_{ij}$  and their first and second derivatives w.r.t.  $\xi$ , evaluated using (5), can be written for the normal noise model (23) as

$$(g_a)_{ij} = \phi_a \sum_{d=1}^D S_{i,d}^T \Sigma_d^{-1} S_{j,d} + E_\xi(h_{ij}), \quad (32)$$

$$\partial_k (g_a)_{ij} = \phi_a \sum_{d=1}^D \left[ (\partial_k S_{i,d})^T \Sigma_d^{-1} S_{j,d} + S_{i,d}^T \Sigma_d^{-1} (\partial_k S_{j,d}) \right] + \partial_k (E_\xi(h_{ij})), \quad (33)$$

where the sensitivity  $S$  is defined as

$$S_{i,d} := \frac{dX_d}{d\xi^i}. \quad (34)$$

Details of the derivation of (32) are given in Appendix C.1. Following [3], we evaluate  $S$  and its partial derivatives for all sampled time points via the numerical solutions to a set of auxiliary differential equations obtained by repeatedly differentiating (20) w.r.t.  $\xi$ . These auxiliary equations are collected in Appendix B.1.

We now explore the application of this methodology with two examples: the Fitzhugh-Nagumo and the Lotka-Volterra model.

*Example: Fitzhugh-Nagumo model.* The Fitzhugh-Nagumo model was developed as a simplification of the Hodgkin-Huxley model, the latter describing the dynamics of the potentials of a spiking neuron. The dynamics of the voltage  $V$  and response  $R$  variables is modelled by the following set of ODEs

$$\frac{dV}{dt} = c \left( V - \frac{V^3}{3} + R \right), \quad (35)$$

$$\frac{dR}{dt} = \frac{a - V - bR}{c}, \quad (36)$$

---

<sup>3</sup>Note: the indices are not summed over in the expression for the log-normal prior.

where  $a, b, c$  are the parameters of the system. Following the treatment in [3] and [13], we simulate the system with starting values  $(V_0, R_0) = (-1, 1)$  and parameters  $(a', b', c') = (0.2, 0.2, 3)$ . Assuming a fixed, identical (over all time-points), normal observation noise model for both potentials with  $\sigma^2 = 0.05$ , we make 25 observations and attempt to infer the parameter values using an SMC sampler with mMALA transitions.

Using  $N = 1000$  particles, a discretisation step-size  $\epsilon = 0.6$ , and a resampling threshold of 0.3, we perform the algorithm over 50 tempered distributions (31) in a geometric sequence (15). We adopt component-wise normal priors for the three parameters. The prior is centred on the simulated mode, i.e.  $(\mu_a, \mu_b, \mu_c) = (0.2, 0.2, 3)$ , and we set the covariance matrix to be  $\Sigma = \text{diag}(0.3^2, 0.3^2, 1.5^2)$ . In addition we impose a positivity constraint on the parameters. The derivatives of  $\dot{V}$  and  $\dot{R}$ , required for the calculation of the Fisher metric, are given in Appendix B.2. We present the scatter plots and weighted histograms representing the full and marginal inferred posterior in Figure 4 and plot the resulting expected ODE solutions in Figure 5. The results verifies the applicability of the SMC sampler to dynamical systems parameter inference.

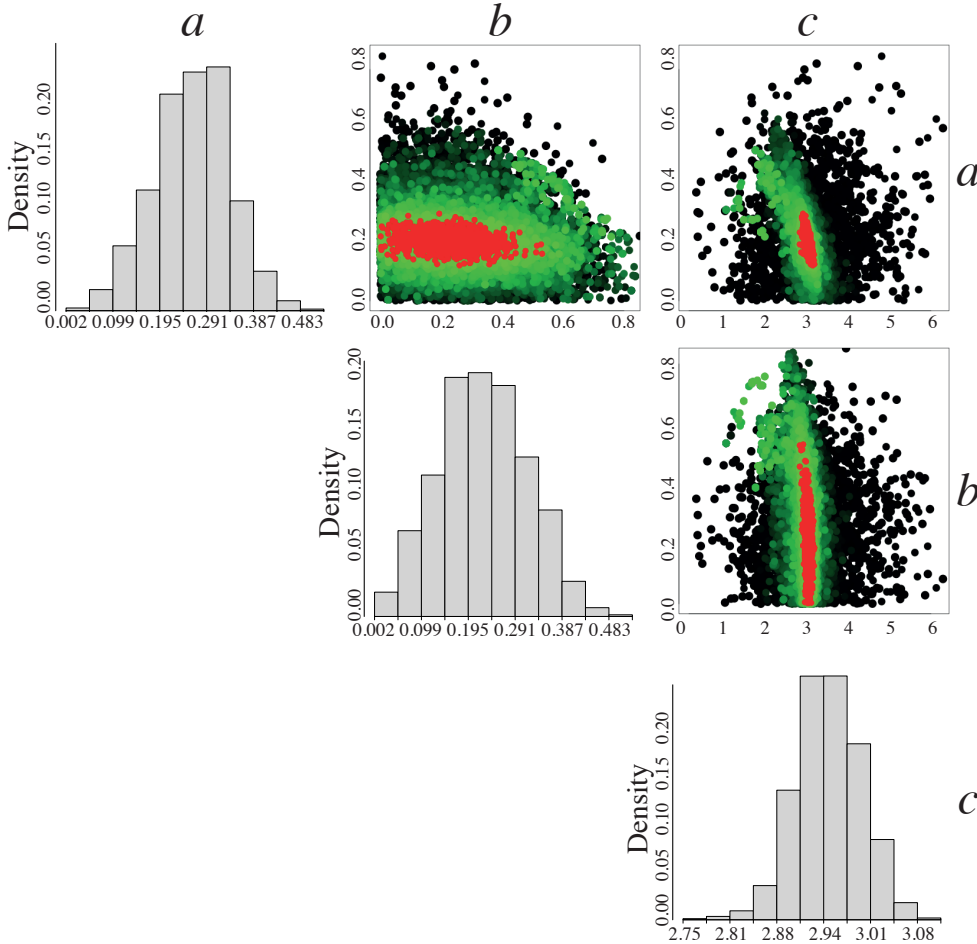


Figure 4: 2D scatterplots and marginal distributions of the three parameters  $a, b, c$  of the Fitzhugh-Nagumo model over 50 populations. The initial particles, coloured black in the scatter plots, are sampled from a component-wise normal prior with a positive truncation. Particles of the intermediate populations are represented in increasingly brighter shades of green with particles of the final population, representing the posterior, coloured red.

We turn now to the issue of the identity of the prior distribution, the importance of which we have already alluded to at various points in our discussion (c.f. (16) and Section 3.3). Similar to the simulations of the static Gaussian example in the previous section, we select an arbitrary grid of points about the simulation mode and focus on the deterministic drift by ignoring the diffusion steps and accepting all particles. We compare the behaviour arising from the component-wise normal prior given above and with that from a component-wise uniform prior with lower and upper limits  $(a_l, b_l, c_l) = (0, 0, 0)$  and  $(a_u, b_u, c_u) = (1, 1, 7)$  respectively. The plots of the particle drifts are given

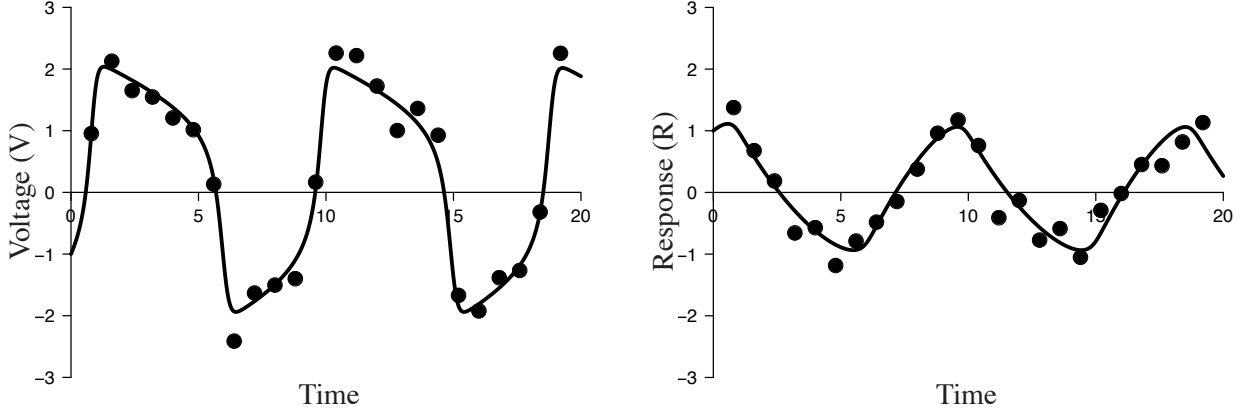


Figure 5: Voltage ( $V$ ) and Response ( $R$ ) observations (dots) and inferred time-series curves (lines). The expected time-series curve is obtained by taking the weighted mean of the inferred curves across all particles from the final population.

in Figure 6. It is clear that, unlike the case with the normal prior, the SMC sampler with a uniform prior does not lead to an orderly drift of particles toward the simulated mode, which results in a highly inefficient sampler where the particle acceptance rate at each MH-step is extremely low ( $< 1\%$ ). This occurrence is unfortunate, if not entirely unexpected for the following reason. The procedure of utilising tempered distributions implies that the components of the Fisher metric (32) corresponding to the initial distributions with low values of the tempering parameter  $\phi_a$  are bounded from below by a function of the prior parameters, i.e. for small  $\phi_a$ ,

$$(g_a)_{ij} \approx E_\xi(h_{ij}). \quad (37)$$

For a uniform prior,  $h_{ij} = 0$  and hence  $(g_a)_{ij} \approx 0$  for all  $i, j$ . Given that both the drift and diffusion terms ((11), (12)) are proportional to the inverse metric  $g_a^{-1}$ , the observed high-temperature drift behaviour<sup>4</sup> in Figure 6 is fully consistent with the theoretical expectations.

In our simulations, it turns out that the SMC sampler with mMALA transitions is not much more robust than one with a non-differential geometric global adaptive kernel, the latter with the computational benefit of not requiring costly computations of the Fisher metric. We speculate on the possible reasons in the next section. However, we now consider another example where the performance of the information geometric SMC sampler shows a clear advantage over a standard adaptive kernel.

*Example: Lotka-Volterra Model.* The Lotka-Volterra model is a simple representation of the predator-prey relationship in a closed environment. Let the variables  $x, y$  represent the numbers of prey and predator species respectively. The dynamics are represented by the set of ODEs

$$\frac{dx}{dt} = x(\alpha - \beta y), \quad (38)$$

$$\frac{dy}{dt} = -y(\gamma - \delta x), \quad (39)$$

where the four parameters  $(\alpha, \beta, \gamma, \delta)$  represent the prey birth rate, prey death rate due to predation, the predator death rate, and the predation efficiency respectively.

We set up a simulation with starting population  $(x_0, y_0) = (15, 30)$  and parameters  $(\alpha', \beta', \gamma', \delta') = (8, 0.5, 0.2, 0.01)$ . We adopt the same assumptions as for the Fitzhugh-Nagumo model simulation but with the normal noise  $\sigma^2 = 0.4$ , and prior parameters centred on the simulated modes with  $\Sigma = \text{diag}(2^2, 0.1^2, 0.05^2, 0.004^2)$ . We let the number of particles  $N = 1000$ , the discretisation step-size  $\epsilon = 0.5$ , and a resampling threshold of 0.3. We perform the algorithm over 30 tempered distributions (31) in a geometric sequence (15). In Figure 7, we show the scatter plots and marginal weighted histograms of the inferred posterior. It is clear from the scatter plots that the parameters are very highly correlated –

<sup>4</sup> $\phi_a \propto \frac{1}{kT}$ , where  $T$  is the temperature.

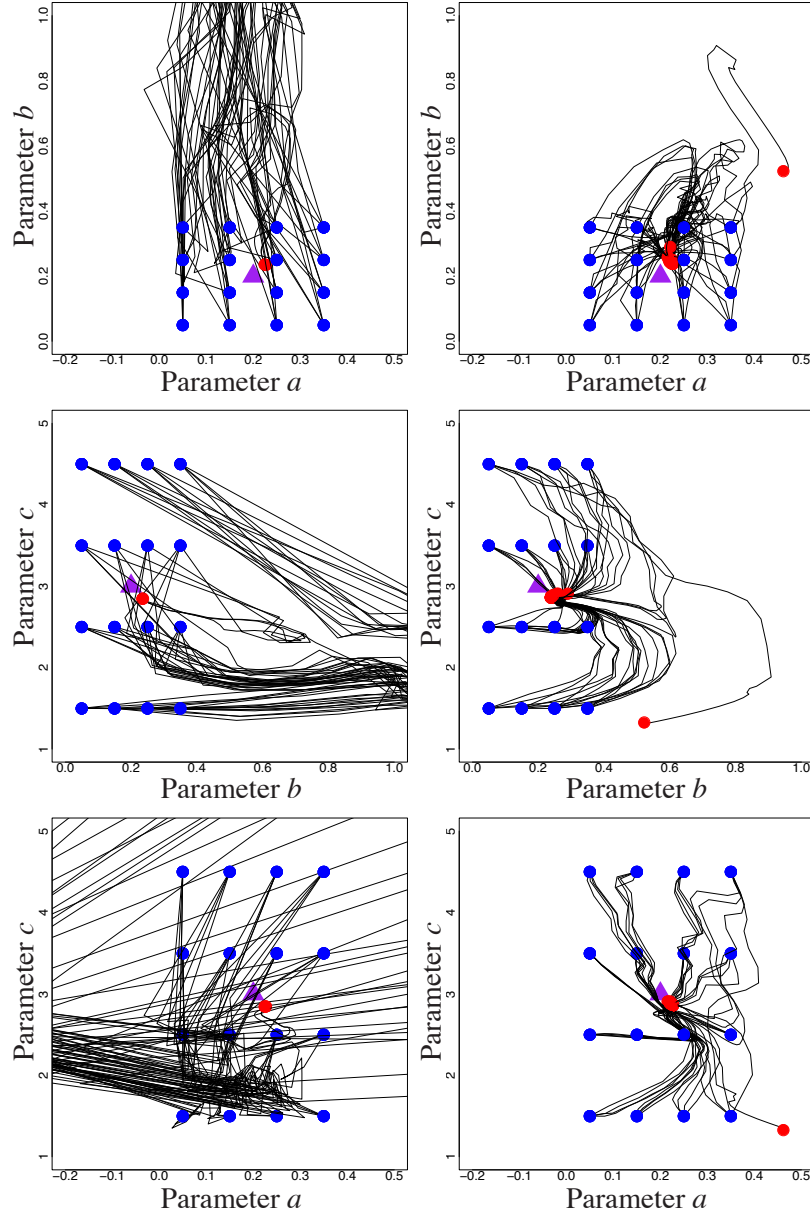


Figure 6: mMALA particle drift paths for the Fitzhugh-Nagumo model. Each of the three rows represent a projection to a separate two-dimensional parameter subspace; the left and right columns are simulations with a uniform and a normal prior respectively. The lack of a lower bound on the Fisher metric in the simulation with the uniform prior (see text) results in a drift behaviour consistent with an extremely high temperature regime, rendering the SMC sampler algorithm unworkable.

precisely the regime where the information geometric approach is expected to confer the greatest benefit. We verify the accuracy of the sampler by simulating the curves with the particle parameters, as shown in Figure 8.

To demonstrate the efficiency of the differential geometric SMC kernel we benchmark its performance with a non-differential geometric, but nevertheless adaptive, kernel. We consider an MCMC kernel with an MVN proposal  $K'(\xi_{a+1} | \xi_a)$  with zero mean and a covariance matrix set to the sample covariance matrix of the particles in population  $a$  multiplied by the asymptotic factor  $(2.38)^2/D$  [14]. Here  $D$  is the dimension of the parameter space (i.e.  $D = 4$ ).

The one measure of efficiency of practical importance we would like to examine is robustness; in the context of parameter inference this means a low variability in the inferred statistics over repeated runs of the SMC sampler. To that end we run, for both the mMALA and the above benchmark kernel, the SMC

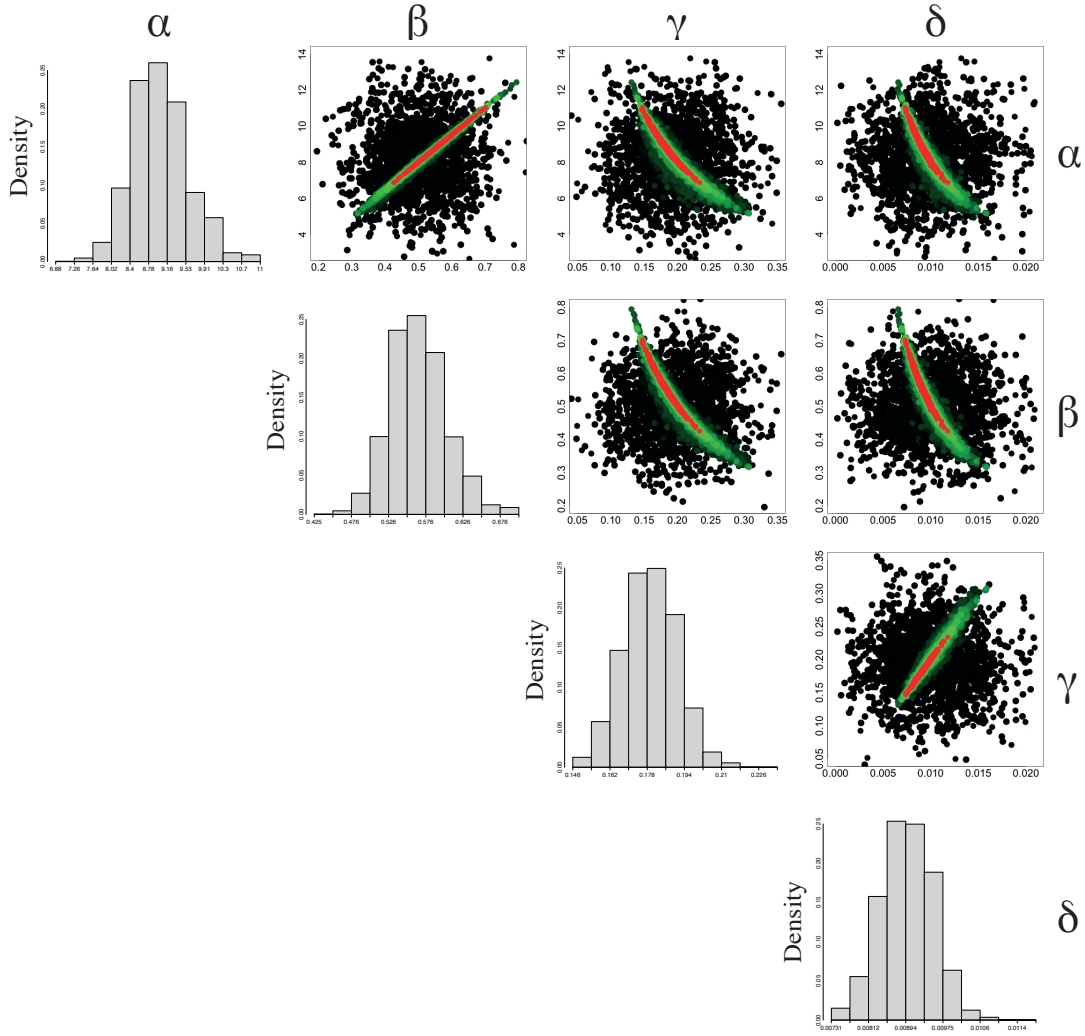


Figure 7: 2D scatterplots and marginal distributions of the four parameters  $\alpha, \beta, \gamma, \delta$  of the Lotka-Volterra model over 30 populations. The initial particles are sampled from a component-wise uniform prior with limits as shown in the figures. These particles are coloured black in the scatter plot. Particles of the intermediate populations are represented in increasingly brighter shades of green with particles of the final population, representing the posterior, coloured red.

sampler for a range of total intermediate populations, performing 27 repetitions of each run. Over each set of 27 runs we determine the sample variance of the inferred parameter means. For both kernels, one would expect the sample variance to decrease with increasing total number population. This is, indeed, what we observe, and the results for parameter  $\alpha$  are given in Figure 9. We see that the mMALA kernel outperforms the non-differential geometric adaptive kernel by demonstrating a consistently high level of robustness over different number of intermediate distributions. By employing the mMALA kernel, we have effectively shortened the distance between successive distributions in the SMC chain, thereby replicating the algorithm with a greater number of intermediate distributions.

An alternative, explanatory, view of the efficiency of the mMALA kernel can be seen by tracking the effective sample sizes and particle acceptance rates of the SMC sampler over the intermediate populations. This is shown in Figure 10. The relatively gradual decline in the ESS of the SMC populations when the mMALA kernel is employed is similar to the observed behaviour ESS along the geodesic in Figure 1.

#### 4. Summary and Discussion

In this paper we have explored the application of methods and ideas from information geometry to SMC sampling and have demonstrated its use in sequential Bayesian inference. We focused on two

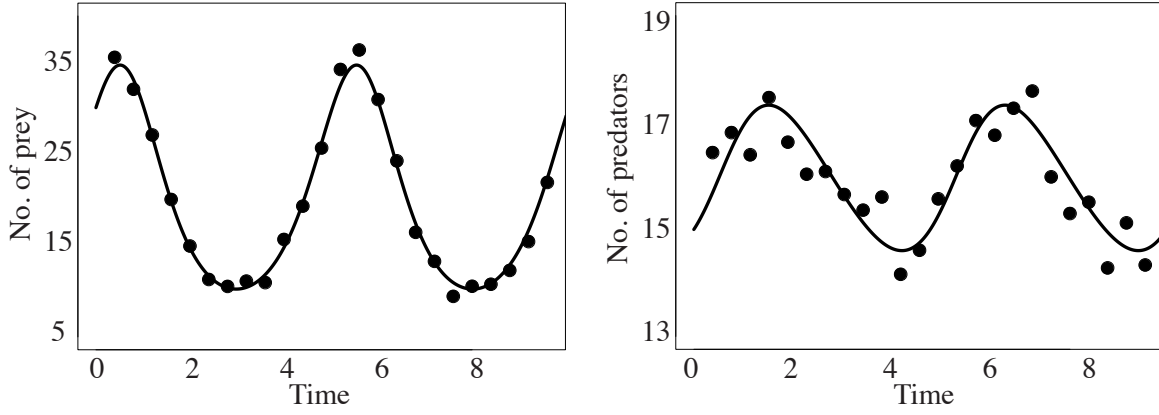


Figure 8: Predator and prey observations (dots) and inferred time-series curves (lines). The expected time-series curves are obtained by taking the weighted mean of the inferred curves across all particles.

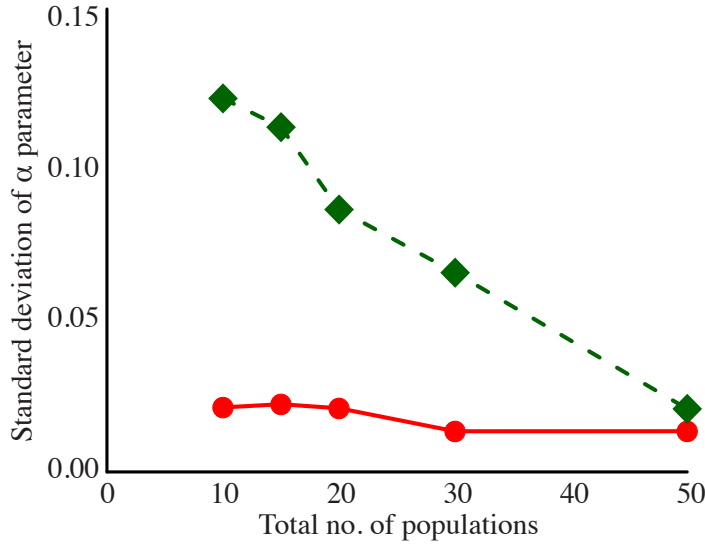


Figure 9: The variance in the estimate of the  $\alpha$  parameter of the Lotka-Volterra model as given by SMC performed using mMALA proposals (**red**  $\bullet$ ) and non-differential geometric global adaptive kernel (**green**  $\blacklozenge$ ). The profile of the estimates of the other three parameters ( $\beta, \gamma, \delta$ ) are similar.

areas – the construction of the sequence of distributions which lie on geodesics, and the employment of mMALA as MCMC transition kernels between the intermediate distributions.

Of the two areas, the theoretical foundations of the mMALA is more established and we have shown that the extension from MCMC to SMC is relatively straightforward. In particular the theoretical differences are the replacement of the single Fisher metric with a sequence of Fisher metrics defined by the sequence of distributions, and the parallelisation of the mMALA diffusion over multiple particles. The conceptual similarity with its use in MCMC allows all the advantages of the original mMALA formulation in Girolami and Calderhead [3] – namely the self-tuning characteristics, efficiency in highly-correlated and high-dimensional settings – to be carried across to SMC.

The issue of intermediate distributions on geodesics is, in comparison, more difficult to tackle, simply due to the lack of analytical solutions for most non-trivial examples. Nevertheless we have demonstrated the potential advantages, in the form of a slower decline in the effective sample sizes, of placing our SMC intermediate distributions on geodesics by way of a simple univariate normal inference example. The computational price of evaluating these geodesics (where available) is a one-off cost incurred at the start of the SMC run and is independent of the number of sampled particles.

At present the computational overhead involved in determining the components of the Fisher metric

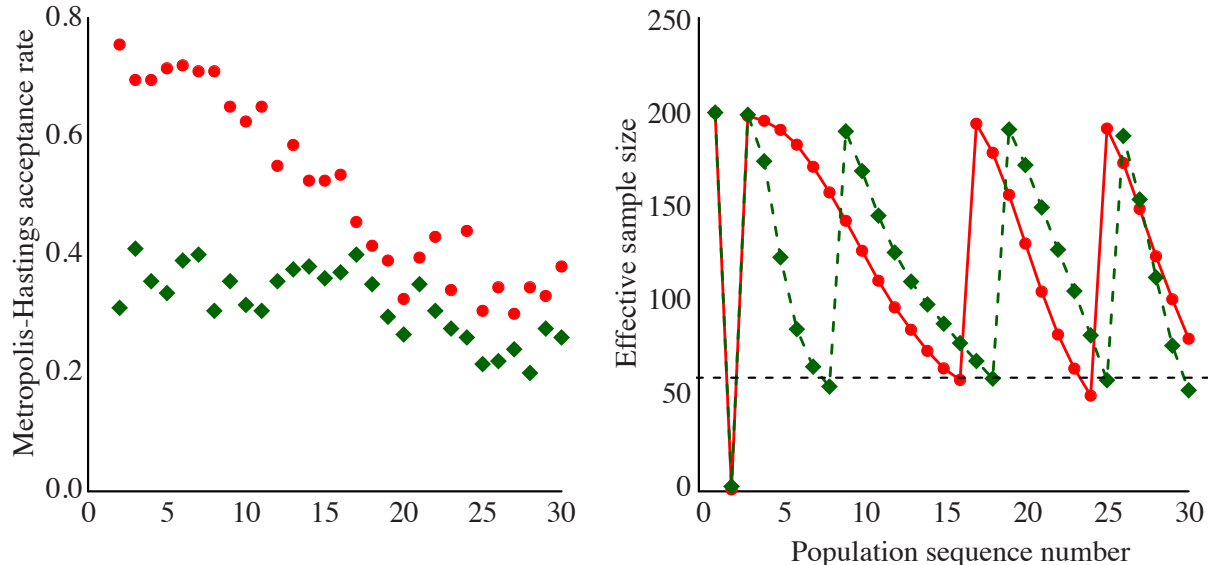


Figure 10: The Metropolis-Hastings acceptance rates (left) and effective sample sizes (right) over a representative single run of the SMC sampler employing the mMALA kernel (red  $\bullet$ ) and the non-differential geometric adaptive random walk kernel (green  $\blacklozenge$ ). The horizontal dotted line indicates the 30% resampling threshold. Employing the mMALA kernel results in a higher MH-acceptance rate and a correspondingly more gradual decline in the ESS.

is seemingly a disadvantage. In the application to the analysis of dynamical systems, one is required to solve at least two separate sets of differential equations for the sensitivities and their derivatives; for heteroscedastic observation noise models, the burden is compounded by the need to solve for the double derivative of the metric. Because the computational cost of evaluating the Fisher metric scales as  $O(D^2)$ , with  $D$  the dimension of the statistical manifold, the efficiency of the technique is reduced in precisely the high-dimensional regime that a non-information geometric kernel might be less suited. For more complicated noise models (e.g. heteroscedasticity), the computational complexity scales as  $O(D^3)$  – this is a consequence of the need to calculate the double derivatives  $\partial_i \partial_j S_k$  (see Appendix C.1). The computational burden can be reduced by adopting the simplified mMALA, where only the first-order sensitivities are calculated [3, 6]. However, as shown in the original MCMC context [3] the effective computational time of the simplified mMALA, expressed in units of the ESS, is not significantly better than its full counterpart. Nevertheless, even in low-dimensional problems, especially in situations where the parameters are highly correlated, as we observed in the Lotka-Volterra model, the advantages conferred by the information geometric approach are clearly evident. Furthermore, it has been suggested<sup>5</sup> that the calculation of the Fisher metric can be made more efficient via a canonical transformation of the coordinates. This is an interesting avenue for further research.

When employing the SMC sampler in sequential Bayesian inference, the form of the prior is often chosen primarily on the basis of knowledge of the parameter constraints, for example a gamma distribution for strictly positive parameters, or a uniform prior when knowledge of the upper and lower bounds is available. Although such bounded priors enforce weak identifiability on the parameter estimation, as was highlighted in [6] and [15], the prior now has the additional role of regularising the Fisher metric. Indeed, as we have shown via the Fitzhugh-Nagumo model example, choosing a uniform prior in conjunction with a tempered sequence of distributions severely impacts the viability of the mMALA kernels; flat or nearly-flat intermediate distributions, as is the case with the initial distributions, results in high-temperature diffusion processes. Because the mMALA diffusion process is informed by the local geometry of the statistical manifold, it does not respect the global density boundaries and attempts to scatter the particles to fill the entire untruncated support, resulting in very low MH-acceptance rates.

Although the transition kernels adapt to the local geometry of the manifold, there is still the out-

<sup>5</sup>By A. C. C. Coolen in the Discussion section of [3].



standing issue of tuning the discretisation step-size parameter. We have shown in the univariate model example that there is a minimum step-size, below which the particles do not travel far enough over the course of the SMC run. One possibility for improvement might be to adopt a higher-order discretisation approximation of the Langevin diffusion SDE which will allow us to select larger step size  $\epsilon$  without deviating too significantly from the Langevin diffusion path, which, in turn, leads to the maintenance of a high MH-acceptance rate and a reduction in the number of intermediate distributions needed for a given level of robustness. This can be achieved by replacing the first-order Euler discretisation with the higher-order Ozaki discretisation [16]. We provide a brief description of the Ozaki discretisation in Appendix C.2. At present it is not clear if the advantages are negated by the added complexity and the resulting increased computational burden, or if the performance of the simplified mMALA relative to the full version can be improved by altering the discretisation scheme.

In summary, we see that the methods of information geometry can be applied to SMC samplers allowing for adaptive and efficient transition kernels. The information theoretic formulation provides a neat and aesthetically pleasing geometrical framework for future improvements in the algorithm. However we have shown that mMALA kernels cannot substitute for careful monitoring of convergence, and potentially adjusting kernels appropriately. This is particularly challenging in areas where likelihoods are flat, which includes many dynamical systems [17, 18, 19].

## Appendix A. Derivations for the geodesic example

### Appendix A.1. Geodesics on the Gaussian distribution manifold

Consider a multivariate normal distribution  $(\mu, \Sigma)$  defining a manifold  $\mathcal{S}$ . Following the analysis in [20, 21, 22], the metric on this space can be written using (5) as

$$ds^2 = (d\mu)^T \Sigma^{-1} d\mu + \frac{1}{2} \text{tr}(\Sigma^{-1} d\Sigma)^2. \quad (\text{A.1})$$

In these coordinates, the geodesic equations (8) are written as

$$\ddot{\Sigma} + \dot{\mu}^T \mu - \dot{\Sigma} \Sigma^{-1} \dot{\Sigma} = 0, \quad (\text{A.2})$$

$$\ddot{\mu} - \dot{\Sigma} \Sigma^{-1} \dot{\mu} = 0. \quad (\text{A.3})$$

The strategy is to first solve for geodesics through the origin  $(\mu(0), \Sigma(0)) = (0, \mathbb{I}_p)$  before translating to the curve with the desired end-points. Adopting the canonical coordinates  $(\Delta, \delta) \equiv (\Sigma^{-1}, \Sigma^{-1}\mu)$ , (A.2) and (A.3) can be partially integrated to give

$$\dot{\Delta} = -B\Delta + x\delta^T, \quad (\text{A.4})$$

$$\dot{\delta} = -B\delta + (1 + \delta^T \Delta^{-1} \delta)x, \quad (\text{A.5})$$

where  $B = \dot{\Delta}(0)$  and  $x = \dot{\delta}(0)$ , and  $\Delta(t)$  and  $\delta(t)$  are the geodesic coordinates parametrised by  $t \in \mathbb{R}$ . It can be shown [22] that the solutions to the geodesic equations are

$$\begin{aligned} \Delta(t) &= \mathbb{I}_p + \frac{1}{2} [\cosh(tG) - \mathbb{I}_p] + \frac{1}{2} B [\cosh(tG) - \mathbb{I}_p] (G^{-1})^2 B \\ &\quad - \frac{1}{2} \sinh(tG) G^{-1} B - \frac{1}{2} B \sinh(tG) G^{-1}, \end{aligned} \quad (\text{A.6})$$

$$\delta(t) = -B [\cosh(tG) - \mathbb{I}_p] (G^{-1})^2 x + \sinh(tG) (G^{-1} x), \quad (\text{A.7})$$

where  $G^2 := B^2 + 2xx^T$ . Simplifying to one dimension ( $p = 1$ ), we have

$$\Delta(t) = 1 + \frac{1}{2} (1 + R^2) (\cosh(tG) - 1) - R \sinh(tG), \quad (\text{A.8})$$

$$\delta(t) = \left( \frac{1 - R^2}{2} \right)^{\frac{1}{2}} (-R (\cosh(tG) - 1) + \sinh(tG)), \quad (\text{A.9})$$

where  $R = B/G$ . This solution would suffice, except that it is given in terms of the gradient terms  $B, G$  instead of the target end-point coordinates; it is however not difficult to rewrite the solution. Keeping ( $p = 1$ ), we set, without loss of generality, the target end-point of the geodesic to be located at  $t = 1$ . Let  $\Delta' \equiv \Delta(1)$  and  $\delta' \equiv \delta(1)$  we solve (A.8) and (A.9) for  $R$  and  $G$  giving

$$R = \frac{\delta'^2 - 2\Delta'^2 + 2\Delta}{(\delta'^4 + 4\delta'^2\Delta'^2 + 4\delta'^2\Delta' + 4\Delta'^4 - 8\Delta'^3 + 4\Delta'^2)^{\frac{1}{2}}}, \quad (\text{A.10})$$

$$G = \cosh^{-1} \left( \frac{\delta'^4 + 4\delta'^2\Delta'^2 + 4\delta'^2\Delta' + 4\Delta'^4 + 4\Delta'^2}{8\Delta'^3} \right). \quad (\text{A.11})$$

Substituting these expressions for  $R$  and  $G$  back into (A.8) and (A.9) gives us the solution for geodesics through the origin and a specified point coordinate at  $t = 1$ .

Now in order to translate geodesics to geodesics, we need a map  $g : \mathcal{S} \rightarrow \mathcal{S}$  which leaves the Fisher metric invariant. As shown in [22] the map which achieves this is the symmetric group  $GA^+(p)/SO(p)$  where the positive affine group is defined as

$$GA^+(p) := \{g = (d, P) \in \mathbb{R}^p \times GL(p, \mathbb{R}) \mid \det P > 0\}, \quad (\text{A.12})$$

and the group action on points in  $\mathcal{S}$ , in terms of coordinates  $(\mu, \Sigma)$  and  $(\Delta, \delta)$ , are

$$\begin{aligned} \mathcal{S} &\longrightarrow \mathcal{S} \\ (\mu, \Sigma) &\longmapsto (P\mu + d, P\Sigma P^T) \\ (\Delta, \delta) &\longmapsto ((P^{-1})^T \Delta P^{-1}, (P^{-1})^T \Delta P^{-1} \Delta^{-1} \delta + \Delta P \Delta^{-1} \delta + \Delta d), \end{aligned} \quad (\text{A.13})$$

and with inverse element  $g^{-1} = (-P^{-1}d, P^{-1})$ .

Given two points  $p_1, p_2 \in \mathcal{S}$ , it is then straightforward to solve for the desired geodesic. First, obtain the group element  $g' \in GA^+(p)/SO(p)$  which maps the origin to  $p_1$ . Substituting  $(\Delta', \delta') = g'^{-1}p_2$  into (A.8) and (A.9) via (A.10) and (A.11), we obtain a solution through the origin which can then be translated using  $g'$  to describe the desired geodesic.

## Appendix A.2. Kernel density for a uniform proposal and Gaussian target

The kernel density in (10) is the Metropolis-Hastings algorithm acceptance probability [23] for a uniform proposal of width  $d$  and a univariate Gaussian target. Following the MH-algorithm, the expression is calculated separately for the cases where the proposal is accepted or rejected. We have

$$K_a(\xi_a \mid \xi_{a-1}) = \begin{cases} 1 - \max \left[ 0, \frac{1}{d\gamma_a(\xi_{a-1})} \left( \Phi \left( -\frac{|\mu - \xi_{a-1}|}{\sigma} \right) - \Phi \left( -\frac{\xi_{a-1} - \mu - a/2}{\sigma} \right) \right) \right] \\ \quad - \max \left[ 0, \frac{1}{d\gamma_a(\xi_{a-1})} \left( \Phi \left( -\frac{\xi_{a-1} - \mu + a/2}{\sigma} \right) - \Phi \left( -\frac{|\mu - \xi_{a-1}|}{\sigma} \right) \right) \right] \\ \quad - \min \left[ \frac{1}{2}, \frac{2}{d} |\mu - \xi_{a-1}| \right], & (\text{for } \xi_a = \xi_{a-1}) \\ \frac{1}{d} \min \left[ 1, \exp \left( -\frac{(\xi_a - \mu)^2}{2\sigma^2} + \frac{(\xi_{a-1} - \mu)^2}{2\sigma^2} \right) \right], & (\text{for } \xi_a \neq \xi_{a-1}) \end{cases} \quad (\text{A.14})$$

with  $(\mu, \sigma)$  the mean and standard deviation parameters of the  $a$ th intermediate distribution.  $\xi_a$  is the particle coordinate of population  $a$  and  $\Phi$  is the cumulative distribution function of the standard normal.

## Appendix B. Non-linear ODE calculations

### Appendix B.1. Auxiliary differential equations for the sensitivities $S_{i,n}$

The subscript conventions used in this section indicates the variable in the differential. We have  $\{i, j, k\} \sim \xi$  and  $\{l, m, n\} \sim X$ , for example  $\partial_j \partial_l f \equiv \frac{\partial^2 f}{\partial \xi_j \partial X^l}$ . To avoid cluttering the notation, when

there is no ambiguity, we drop the  $X$ -index on  $\dot{S}$ , i.e.  $\dot{S}_{i,l} \rightarrow \dot{S}_i$  and  $f_l \rightarrow f$ .

$$\dot{S}_i = (\partial_l f) S_{i,l} + \partial_i f, \quad (\text{B.1})$$

$$\partial_k \dot{S}_i = (\partial_m \partial_l f) S_{k,m} S_{i,l} + (\partial_k \partial_l f) S_{i,l} + (\partial_l f) (\partial_k S_{i,l}) + (\partial_l \partial_i f) S_{k,l} + \partial_i \partial_k f, \quad (\text{B.2})$$

$$\begin{aligned} \partial_j \partial_k \dot{S}_i &= (\partial_m \partial_l \partial_j f) S_{k,m} S_{i,l} + (\partial_m \partial_l f) (\partial_j S_{k,m}) S_{i,l} + (\partial_m \partial_l f) (\partial_j S_{i,l}) S_{k,m} \\ &\quad + (\partial_k \partial_j \partial_l f) S_{i,l} + (\partial_k \partial_l f) (\partial_j S_{i,l}) + (\partial_l \partial_j f) (\partial_k S_{i,l}) + (\partial_l f) (\partial_j \partial_k S_{i,l}) \\ &\quad + (\partial_l \partial_j \partial_i f) S_{k,l} + (\partial_l \partial_i f) (\partial_j S_{k,l}) + \partial_i \partial_j \partial_k f \\ &\quad + S_{j,n} [(\partial_m \partial_l \partial_n f) S_{k,m} S_{i,l} + (\partial_k \partial_l \partial_n f) S_{i,l} + (\partial_l \partial_n f) (\partial_k S_{i,l}) \\ &\quad + (\partial_l \partial_i \partial_n f) S_{k,l} + \partial_i \partial_k \partial_n f]. \end{aligned} \quad (\text{B.3})$$

## Appendix B.2. Partial derivatives for the ODE examples

*Fitzhugh-Nagumo model.* We reproduce here the expressions in [3]. The two components are labelled  $(V, R)$  with parameters  $(a, b, c)$ . The non-zero single derivatives are

$$\begin{aligned} \frac{\partial \dot{V}}{\partial c} &= V - \frac{V^3}{3} + R, \quad \frac{\partial \dot{R}}{\partial a} = \frac{1}{c}, \quad \frac{\partial \dot{R}}{\partial b} = -\frac{R}{c}, \quad \frac{\partial \dot{R}}{\partial c} = \frac{V - a + bR}{c^2}, \\ \frac{\partial \dot{V}}{\partial V} &= c(1 - V^2), \quad \frac{\partial \dot{V}}{\partial R} = c, \quad \frac{\partial \dot{R}}{\partial V} = -\frac{1}{c}, \quad \frac{\partial \dot{R}}{\partial R} = -\frac{b}{c}, \end{aligned} \quad (\text{B.4})$$

and the double derivatives

$$\begin{aligned} \frac{\partial^2 \dot{R}}{\partial^2 c} &= \frac{2(a - V - bR)}{c^3}, \quad \frac{\partial^2 \dot{R}}{\partial a \partial c} = -\frac{1}{c^2}, \quad \frac{\partial^2 \dot{R}}{\partial b \partial c} = \frac{R}{c^2}, \\ \frac{\partial^2 \dot{V}}{\partial V \partial c} &= 1 - V^2, \quad \frac{\partial^2 \dot{V}}{\partial R \partial c} = 1, \quad \frac{\partial^2 \dot{R}}{\partial R \partial b} = -\frac{1}{c}, \quad \frac{\partial^2 \dot{R}}{\partial R \partial c} = \frac{b}{c^2}. \end{aligned} \quad (\text{B.5})$$

*Lotka-Volterra model.* Let  $(x, y)$  label the prey and predator respectively with parameters  $(\alpha, \beta, \gamma, \delta)$ . The non-zero single derivatives are

$$\begin{aligned} \frac{\partial \dot{x}}{\partial \alpha} &= x, \quad \frac{\partial \dot{x}}{\partial \beta} = -xy, \quad \frac{\partial \dot{y}}{\partial \gamma} = -y, \quad \frac{\partial \dot{y}}{\partial \delta} = xy, \\ \frac{\partial \dot{x}}{\partial x} &= \alpha - \beta y, \quad \frac{\partial \dot{x}}{\partial y} = -\beta x, \quad \frac{\partial \dot{y}}{\partial x} = \delta y, \quad \frac{\partial \dot{y}}{\partial y} = -(\gamma - \delta x), \end{aligned} \quad (\text{B.6})$$

and the double derivatives are

$$\begin{aligned} \frac{\partial^2 \dot{x}}{\partial x \partial \alpha} &= -\frac{\partial^2 \dot{y}}{\partial y \partial \gamma} = 1, \quad \frac{\partial^2 \dot{x}}{\partial x \partial \beta} = -\frac{\partial^2 \dot{y}}{\partial x \partial \delta} = -y, \quad \frac{\partial^2 \dot{x}}{\partial y \partial \beta} = -\frac{\partial^2 \dot{y}}{\partial y \partial \delta} = -x, \\ \frac{\partial^2 \dot{x}}{\partial x \partial y} &= -\beta, \quad \frac{\partial^2 \dot{y}}{\partial x \partial y} = \delta. \end{aligned} \quad (\text{B.7})$$

## Appendix C. Extensions of mMALA

### Appendix C.1. Heteroscedasticity and bounded parameters

In this section we present the implications on the calculation of the Fisher metric of incorporating additional structure on top of the normal noise model. In particular, we focus on heteroscedasticity and boundary constraints in observation noise. In the first case, in addition to a fixed normal observation noise, we have a noise contribution which scale scales with  $X$ ; here the noise model is given by (23) but with the replacement

$$\Sigma_d := \mathbb{I}_\tau \sigma_d^2 + \text{diag}((\mathbb{I}_\tau \tilde{\sigma}_d^2)(X_{\cdot,d} * X_{\cdot,d})). \quad (\text{C.1})$$

In the second case, because  $Y$  often represents measurement of physical quantities (concentration, volume, etc) there is usually a positive constraint on  $X$  and  $Y$ , the implication being that one has to consider truncated distributions.

Using the expression of  $\Phi$  from (23) for the multivariate normal and a simple application of the chain rule, we have

$$\begin{aligned} -\partial_i \partial_j \Phi_d &= \partial_i \partial_j \left( \frac{1}{2} (Y_d - X_d)^T \Sigma_d^{-1} (Y_d - X_d) \right) \\ &= S_{d,i}^T \Sigma_d^{-1} S_{d,j} - (Y_d - X_d)^T (\Sigma_d^{-1} (\partial_i S_{d,j}) + (\partial_i \Sigma_d^{-1}) S_{d,j} + (\partial_j \Sigma_d^{-1}) S_{d,i}) \\ &\quad + \frac{1}{2} (Y_d - X_d)^T (\partial_i \partial_j \Sigma_d^{-1}) (Y_d - X_d). \end{aligned} \quad (\text{C.2})$$

For a normal distribution truncated between limits  $a < X < b$ , the expectation and variance can be written in terms of the untruncated mean and s.d.  $(\mu, \sigma)$  as [24]

$$E(X|a < X < b) = \mu + \frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \sigma, \quad (\text{C.3})$$

$$\text{Var}(X|a < X < b) = \sigma^2 \left[ 1 + \frac{\frac{a-\mu}{\sigma} \phi(\frac{a-\mu}{\sigma}) - \frac{b-\mu}{\sigma} \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} - \left( \frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{a-\mu}{\sigma}) - \Phi(\frac{b-\mu}{\sigma})} \right)^2 \right]. \quad (\text{C.4})$$

We evaluate the Fisher metric as  $[g_a(\xi)]_{ij} = -\phi_a E_\xi(\partial_i \partial_j \Phi - h_{ij})$ , where  $h_{ij}$  is given by (28)–(30). Without loss of generality, we set  $h_{ij} = 0$  and the Fisher metrics  $(g_a)_{ij}$  and their first and second derivatives w.r.t.  $\xi$ , evaluated using (5), can be written for the noise model (23) with both the heteroscedasticity and positivity constraint as

$$(g_a)_{ij} = \phi_a \sum_{d=1}^D \left[ S_{i,d}^T \Sigma_d^{-1} S_{j,d} - (L_{d,ij}(\lambda(\alpha_d) * \sqrt{K_d})) + \frac{1}{2} \text{tr}((\partial_i \partial_j \Sigma_d^{-1}) J_d) \right], \quad (\text{C.5})$$

$$\begin{aligned} \partial_k (g_a)_{ij} &= \phi_a \sum_{d=1}^D \left[ (\partial_k S_{i,d})^T \Sigma_d^{-1} S_{j,d} + S_{i,d}^T \Sigma_d^{-1} (\partial_k S_{j,d}) \right. \\ &\quad + S_{d,i}^T (\partial_k \Sigma_d^{-1}) S_{d,j} - (\partial_k (\lambda(\alpha_d) * \sqrt{K_d})^T) (\Sigma_d^{-1} (\partial_i S_{d,j}) + (\partial_i \Sigma_d^{-1}) S_{d,j} \\ &\quad + (\partial_j \Sigma_d^{-1}) S_{d,i}) - (\lambda(\alpha_d) * \sqrt{K_d})^T [(\partial_k \Sigma_d^{-1}) (\partial_i S_{d,i}) + \Sigma_d^{-1} (\partial_k \partial_i S_{d,i}) \\ &\quad + (\partial_k \partial_i \Sigma_d^{-1})^T S_{d,j} + (\partial_i \Sigma_d^{-1} (\partial_k S_{d,k}) + (\partial_k \partial_j \Sigma_d^{-1}) S_{d,i} + (\partial_j \Sigma_d^{-1}) (\partial_k S_{d,i}))] \\ &\quad \left. + \frac{1}{2} \text{tr}((\partial_k \partial_j \partial_i \Sigma_d^{-1})^T J_d) + (\partial_i \partial_j \Sigma_d^{-1})^T \partial_d J_d \right), \end{aligned} \quad (\text{C.6})$$

where

$$L_{d,ij} := \Sigma_d^{-1} (\partial_i S_{d,j}) + (\partial_i \Sigma_d^{-1}) S_{d,j} + (\partial_j \Sigma_d^{-1}) S_{d,i}, \quad (\text{C.7})$$

$$J_d := K_d * (\mathbf{1}_p - \alpha_d * \lambda(\alpha_d)). \quad (\text{C.8})$$

Specifying the fixed and heteroscedastic noise components  $\sigma$  and  $\tilde{\sigma}$ , the three terms  $K_d$ ,  $\lambda(\alpha_d)$  and  $\alpha_d$  are defined by

$$\Sigma_d = \text{diag}(K_d), \quad \alpha_d := X_d * K_d^{-1}, \quad [\lambda(\alpha_d)]_i := \frac{\phi(\alpha_{d,i})}{\Phi(\alpha_{d,i})}, \quad (\text{C.9})$$

with  $\Sigma_d$  defined in (C.1),  $X_d \equiv X_{\cdot,d}$ , i.e. the  $d^{\text{th}}$  column of  $X$ , and  $\phi(\cdot)$  and  $\Phi(\cdot)$  the density and cumulative distribution functions respectively of the standard normal<sup>6</sup>.

Removing the positivity constraints and setting the heteroscedastic variance term to zero, i.e.  $a \rightarrow$

---

<sup>6</sup>Not to be confused with the tempering parameter  $\phi_a$  or the exponent  $\Phi$  in the likelihood in (27).

$-\infty$  and  $\tilde{\sigma}_d = 0$ , we recover the expressions of the Fisher metric and its derivatives in the main text (32) and (33), respectively.

### Appendix C.2. Higher-order Ozaki discretisation of the Langevin diffusion

The Ozaki discretisation was formulated by Ozaki and Shoji [16, 25] and proposed for use in MALA framework by Stramer and Tweedie [26]. In contrast to the Euler discretisation, it is more stable [27], providing a higher-order approximation for the drift term in the Langevin diffusion SDE. On the downside, its implementation can often be computationally expensive.

The Ozaki discretisation of the Langevin diffusion SDE can be expressed as an MVN proposal, like the Euler discretisation in (11), as

$$q_a(\xi_{a+1} \mid \xi_a) \sim \mathcal{N}(\mu_a^O(\xi_a, \epsilon), \Sigma_a^O(\xi_a, \epsilon)), \quad (\text{C.10})$$

where the components of the sequence (as indexed by  $a$ ) of means  $\mu_a^O(\xi_a, \epsilon)$  and covariance matrices  $\Sigma_a^O(\xi_a, \epsilon)$  are given by

$$(\mu_a^O(\xi_a, \epsilon))^i = (\xi_a)^i + \left( J_a^{-1}(\exp(\epsilon^2 J_a) - \mathbb{I}_D) \right)_j^i b^j, \quad (\text{C.11})$$

$$(\Sigma_a^O(\xi_a, \epsilon))^{ij} = \frac{1}{2} g^{il} \left( J_a^{-1}(\exp(2\epsilon^2 J_a) - \mathbb{I}_D) \right)_l^j, \quad (\text{C.12})$$

where the drift term  $b_a(\xi_a)$  is written as

$$b_a(\xi_a) = \frac{1}{2} g^{ij} \partial_j \ell - g^{ik} (\partial_j g_{kl}) g^{lj} + \frac{1}{2} g^{ij} g^{kl} \partial_j g_{kl}, \quad (\text{C.13})$$

with the Jacobian  $(J_a(\xi_a))_j^i = \frac{\partial b_a^i}{\partial \xi_a^j}$ . The explicit form of  $J_a(\xi_a)$  is given by

$$\begin{aligned} J_a(\xi_a) = & -\frac{1}{2} (g^{ik} (\partial_j g_{kl}) g^{lm} \partial_m \ell + \frac{1}{2} \partial_j \partial_k \ell + g^{ik} (\partial_j g_{kl}) g^{lm} (\partial_n g_{mp}) g^{pn} \\ & - g^{ik} (\partial_j \partial_l g_{km}) g^{ml} + g^{ik} (\partial_n g_{kl}) g^{lm} (\partial_j g_{mp}) g^{pn} - \frac{1}{2} g^{ik} (\partial_j g_{kl}) g^{lm} g^{np} (\partial_m g_{pn}) \\ & + g^{ik} [g^{lm} (\partial_j g_{mn}) g^{np} (\partial_k g_{pl}) + g^{lm} (\partial_j \partial_k g_{lm})]) \end{aligned} \quad (\text{C.14})$$

Expanding the first two terms of the exponentials in (C.11) and (C.12), we have

$$(\mu_a^O(\xi_a, \epsilon))^i \approx \xi_a^i + \epsilon^2 b_a^i + \frac{\epsilon^4}{2} (J_a b_a)^i, \quad (\text{C.15})$$

$$(\Sigma_a^O(\xi_a, \epsilon))^{ij} \approx g^{il} (\epsilon^2 \mathbb{I}_D + 2\epsilon^4 J_a)_l^j. \quad (\text{C.16})$$

Keeping only the  $O(\epsilon^2)$  terms of the expansion, we recover the Euler discretisation proposal parameters (13).

### Acknowledgements

AS and MPHS gratefully acknowledge financial support from EPSRC (EP/I017267/1); SF is funded through an MRC Computational Biology Research Fellowship; MPHS is a Royal Society Research Merit award holder.

### References

- [1] A. Doucet, N. de Freitas, N. Gordon, Sequential Monte Carlo Methods in Practice, Springer, 2010.

- [2] P. Del Moral, A. Doucet, A. Jasra, Sequential Monte Carlo samplers, *Journal of the Royal Statistical Society. Series B. Statistical Methodology* 68 (2006) 411–436.
- [3] M. Girolami, B. Calderhead, Riemann manifold Langevin and Hamiltonian Monte Carlo methods, *Journal of the Royal Statistical Society. Series B. Statistical Methodology* 73 (2011) 123–214.
- [4] A. Honkela, T. Raiko, M. Kuusela, M. Törnio, J. Karhunen, Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes, *Journal of Machine Learning Research (JMLR)* 11 (2010) 3235–3268.
- [5] M. Komorowski, M. J. Costa, D. A. Rand, M. P. H. Stumpf, Sensitivity, robustness, and identifiability in stochastic chemical kinetics models., *Proceedings of the National Academy of Sciences of the United States of America* 108 (2011) 8645–8650.
- [6] B. Calderhead, M. Girolami, Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods, *Interface Focus* (2011).
- [7] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, 2008.
- [8] S.-I. Amari, H. Nagaoka, *Methods of Information Geometry* (Translations of Mathematical Monographs) (Translations of Mathematical Monographs), American Mathematical Society, 2007.
- [9] S. Kullback, *Information Theory and Statistics*, Courier Dover Publications, 1968.
- [10] S. I. R. Costa, S. A. Santos, J. E. Strapasson, Fisher information distance: a geometrical reading?, *arXiv:1210.2354 [stat.ME]* (2012).
- [11] R. M. Neal, Sampling from multimodal distributions using tempered transitions, *Statistics and computing* 6 (1996) 353–366.
- [12] G. Behrens, N. Friel, M. Hurn, Tuning tempered transitions, *Statistics and computing* 22 (2012) 65–78.
- [13] J. O. Ramsay, G. Hooker, D. Campbell, J. Cao, Parameter estimation for differential equations: a generalized smoothing approach, *Journal of the Royal Statistical Society. Series B. Statistical Methodology* 69 (2007) 741–796.
- [14] Y. Atchadé, G. Fort, E. Moulines, P. Priouret, Adaptive Markov chain Monte Carlo: theory and methods, in: *Bayesian Time Series Models*, Cambridge Univ. Press, Cambridge, 2011, pp. 32–51.
- [15] M. Vallisneri, Use and abuse of the Fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects, *Physical Review D* 77 (2008) 042001.
- [16] T. Ozaki, A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: a local linearization approach, *Statistica Sinica* (1992).
- [17] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, J. P. Sethna, Universally sloppy parameter sensitivities in systems biology models, *PLoS Comput Biol* 3 (2007) 1871–1878.
- [18] J. F. Apgar, D. K. Witmer, F. M. White, B. Tidor, Sloppy models, parameter uncertainty, and the role of experimental design, *Molecular Biosystems* (2010).
- [19] K. Erguler, M. P. H. Stumpf, Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models., *Molecular Biosystems* 7 (2011) 1593–1602.
- [20] M. Calvo, J. M. Oller, An explicit solution of information geodesic equations for the multivariate normal model, *Statistics & Decisions. International Journal for Statistical Theory and Related Fields* 9 (1991) 119–138.
- [21] L. T. Skovgaard, A Riemannian Geometry of the Multivariate Normal Model, *Scandinavian journal of statistics* 11 (1984) 211–223.
- [22] T. Imai, A. Takaesu, M. Wakayama, Remarks on geodesics for multivariate normal models, 2011B-6 (2011).

- [23] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of State Calculations by Fast Computing Machines, *The Journal of Chemical Physics* 21 (1953) 1087.
- [24] N. L. Johnson, S. Kotz, N. Balakrishnan, *Continuous univariate distributions*, Wiley-Interscience, 1995.
- [25] I. Shoji, Miscellanea. A statistical method of estimation and simulation for systems of stochastic differential equations, *Biometrika* 85 (1998) 240–243.
- [26] O. Stramer, R. L. Tweedie, Langevin-Type Models I: Diffusions with Given Stationary Distributions and their Discretizations, *Methodology and Computing in Applied Probability* 1 (1999) 283–306.
- [27] G. O. Roberts, O. Stramer, Langevin diffusions and Metropolis-Hastings algorithms, *Methodology and Computing in Applied Probability* 4 (2002) 337–357.